LAMP-TR-078 September 2001
CS-TR-4285
UMIACS-TR-2001-64

Spanish Language Processing at University of Maryland: Building Infrastructure for Multilingual Applications

Clare Cabezas, Bonnie Dorr, Philip Resnik

Language and Media Processing Labratory
Instititue for Advanced Computer Studies
College Park, MD 20742

Abstract

We describe here our construction of lexical resources, tool creation, building of an aligned parallel corpus, and an approach to automatic treebank creation that we have been developing using Spanish data, based on projection of English syntactic dependency information across a parallel corpus.

***The support of the LAMP Technical Report Series and the partial support of this research by the National Science Foundation under grant EIA0130422 and the Department of Defense under contract MDA9049-C6-1250 is gratefully acknowledged.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
SEP 2001		**00-09-2001 to 00-09-2001**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Spanish Language Processing at University of Maryland: Building Infrastructure for Multilingual Applications	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**6**	

Spanish Language Processing at University of Maryland: Building Infrastructure for Multilingual Applications

Clara Cabezas, Bonnie Dorr, Philip Resnik
University of Maryland, College Park, MD 20742
{clarac,bonnie,resnik}@umiacs.umd.edu

Abstract: We describe here our construction of lexical resources, tool creation, building of an aligned parallel corpus, and an approach to automatic treebank creation that we have been developing using Spanish data, based on projection of English syntactic dependency information across a parallel corpus.

Introduction

NLP researchers at the University of Maryland are currently working on the construction of resources and tools for several multilingual applications, with a focus on broad coverage machine translation (MT) and cross-language information retrieval. We describe here our construction of lexical resources, tool creation, building of an aligned parallel corpus, and an approach to automatic treebank creation, which we have been developing using Spanish data, based on projection of English syntactic dependency information across a parallel corpus.

Creating lexical databases for Spanish

We have built two types of lexical databases for Spanish: one that is semantico-syntactic, based on a representation called *Lexical Conceptual Structure* (LCS), and one that is morphological, based on Kimmo-style Spanish entries.

An LCS is a directed graph with a root that reflects the semantics of a lexical item by a combination of semantic structure and semantics content. LCS representations are both language and structure independent; they were originally formulated by Jackendoff (1983, 1990) and have been used as interlingua in a number of machine translation projects including UNITRAN and MILT (Dorr 1993; Dorr 1997).

The creation of a Spanish LCS lexicon relied heavily on the existence of a large hand-generated database of English LCS entries, which were ported over to Spanish LCS entries by means of a bilingual lexicon and acquisition procedures as described in Dorr (1997).

Our Spanish morphological lexicon was originally derived from a two-level Kimmo-based morphology system (Dorr 1993). This lexicon contains 273 roots and 99 types of endings, with an upper bound of possible morphological realizations of 27,027 (the product of number of roots, multiplied by the number of endings).

The structure of the lexicon consists of (a) root word entries, (b) continuation classes, and (c) endings.

```
(DEF-MORPH-ROOT language root features
   (string-root1 continuation-class1 features1)
   (string-root2 continuation-class2 features2))
```

For example, the Spanish words ‘veo’ (‘I see’) and ‘visto’ (‘seen’), would have the following entries in the lexicon:

```
(DEF-MORPH-ROOT Spanish VER [v]
   (“ve” *ER-IRREG-6 NIL)
   (“visto” NIL [perf-tns]))
```

We used this lexicon for English-to-Spanish query translation in several cross-language information retrieval experiments. The results were presented at the First International Conference on Language Resource Evaluation (LREC) in Granada, Spain (Dorr and Oard 1998).

Applying a Spanish LCS Lexicon in MT

We have experimented with an interlingual approach to Spanish-English machine translation, using LCS representations as the interlingua. In our most recent experiments in Spanish to English translation, we have used LCS together with Abstract Meaning Representations (AMR) as developed at USC/ISI (Langkilde and Knight, 1998a). AMRs are semantic-syntactic language-specific representations.

After parsing the Spanish sentence, we create a semantic representation (LCS), which is then transformed into a syntactic-semantic representation of the target language sentence (AMR). This representation serves as the input to

Nitrogen, a generation tool developed by USC/ISI (Langkilde and Knight, 1998a; Langkilde and Knight 1998b). Nitrogen is responsible for (a) transforming the Spanish syntactic representation into an English syntactic representation, (b) Creating a word by generating all the possible surface orderings (linearizations) for the English sentence, (c) Using a n-gram language model to choose the optimal linearization, and finally (d) generating morphological realizations, i.e. producing the surface form for the English sentence which corresponds to the translation of the Spanish original sentence.

Acquiring bilingual dictionary entries

In addition to building and applying the more sophisticated LCS lexical representations, we have explored the automatic acquisition of simple word-to-word correspondences from parallel corpora, based on cross-language statistical association between word co-occurrences. The noisy, confidence-ranked bilingual lexicons obtained in this way can be useful in porting LCS lexicons to new languages, as described above, and are also useful by themselves in improving dictionary-based cross language information retrieval (Resnik, Oard, and Levow, 2001).

Constructing an Aligned Corpus

Parallel corpora have emerged as a crucial resource for acquiring and improving lexical resources such as bilingual lexicons, and for developing broad coverage machine translation techniques. We have therefore devoted effort to acquiring English-Spanish parallel text using traditional and less traditional channels.

Collecting Parallel Text

We have obtained parallel data in three ways. First, we have taken advantage of community-wide corpus distribution channels, such as the Linguistic Data Consortium (LDC), the European Language Resource Distribution Agency (ELDA) and the Foreign Broadcast Information Service (FBIS). These sources provide data that are generally clean and often aligned or easily alignable, and which have the advantage of being available in common to a large community of researchers.

Second, we have collected parallel text from the World Wide Web using the STRAND system for acquiring parallel Web documents (Resnik, 1999). (One such collection of Spanish-English documents is available, as a set of URL pairs, at http://umiacs.umd.edu/~resnik/strand.) Data collected from the Web have the advantage of great diversity in contrast to the often more domain- or genre-specific forms of text available from standard sources; on the other hand, they are also often of extremely diverse quality.

Third, we have obtained a parallel English-Spanish version of the Bible as part of our general project collecting freely available Bible versions and annotating their parallel structure using the Corpus Encoding Standard (CES), as a parallel resource for use in computational linguistics. Our empirical studies of the Bible's size and vocabulary coverage – using LDOCE and the Brown Corpus for comparison – suggest that modern-language Bibles are a surprisingly viable source of information about everyday language research (Resnik, Olsen, and Diab, 1999). CES-annotated parallel English and Spanish versions are available on the Web at http://umiacs.umd.edu/~resnik/parallel/.

In the work we describe here, we have been focusing our development on the Spanish-English United Nations Parallel Corpus, available from LDC, which has data generated from 1989 through 1991.

Aligning the Text at the Sentence Level

The U.N. Parallel Corpus is already aligned at the document level. Our alignment of the corpus at lower levels uses a combination of existing tools and components we have constructed.

As a first stage in below-document-level alignment, we preprocess the text in order to obtain alignments at the paragraph level using simple document structure. HTML-style markup, indicating a number of within-text boundaries above the sentence level, is introduced automatically on the basis of relevant cues in the text. The resulting marked-up document is passed to a structure-based alignment tool designed for use with HTML documents (Resnik, 1999), which uses dynamic programming (Unix *diff*) to generate an alignment between text chunks on the basis of correspondences

in markup. Because only boundary markup is used, not content, the process is entirely language independent. Although the introduction of markup is pattern-based and therefore somewhat heuristic, it succeeds well at avoiding the introduction of spurious (intra-sentential) boundaries.

Next, we used MXTERMINATOR (Reynar and Ratnaparkhi, 1997) to break multi-sentence chunks into sentences boundaries both in Spanish and English. This is a supervised system based on maximum entropy models that learns sentence boundaries from correctly boundary-annotated text. Thus far we have used a version trained on English text, and we have found that it performs reasonably well for both Spanish and English. Our sentence-level alignment of the U.N. parallel data produced roughly 300,000 sentences per side.

Tokenization

Our ultimate goal being word-level alignment, we required tokenized text. We implemented a tokenizer for Spanish using a number of Perl pattern matching rules, some of them adapted from the Spanish Kimmo-style morphological analyzer (Dorr, 1993). In its current state, this tokenizer removes SGML tags, bad spacing characters (tabs/spaces/ansi space/ etc.) and punctuation (in the case of periods at the end of the sentence, it actually separates them from the preceding word). It also merges over 2000 frequently co-occurring words that form fixed expressions, e.g. the tokens in 'dentro de' will be merged into 'dentro_de'. Finally, it performs morphological analysis. In the case of verbs, it uses 70 Perl substitution rules in order to make sure that the accentuation patterns and spelling change according to the resulting verb base form. For example, the first person singular 'finjo' (I fake) becomes the infinitive 'fingir' and not *'finjir'. This tokenizer has been used in our initial dependency tree inference experiments for Spanish, described below.

Aligning Text at the Word Level

Once the text has been reduced to aligned sentences, we train IBM statistical MT models using software developed by Al-Onaizan et al. (1999). The training process produces model parameters and, as a side-effect, it produces the most likely word-level alignment for each sentence pair in the training corpus. Preliminary analysis of these alignments is what led us to move from an extremely unsophisticated Spanish tokenizer to one that takes into account morphology and frequent multi-word co-occurrences.

Creating a Noisy Spanish Treebank

Statistical methods in NLP have led to major advances, with supervised training methods leading the way to the greatest improvements in performance on tasks such as part-of-speech tagging, syntactic disambiguation, and broad-coverage parsing. Unfortunately, the annotated data needed for supervised training are available for only a small number of languages.

The University of Maryland has recently begun a project in collaboration with Johns Hopkins University aimed at breaking past this bottleneck. A central idea in this effort is to take advantage of the rich resources available for English, together with parallel corpora: the English side of a parallel corpus is annotated using existing tools and resources, and the results projected to the language on the other side using word-level alignments as a bridge; finally supervised training is used to create tools that perform well despite noise in the automatically annotated corpus. Yarowsky *et al.* (2001) have shown extremely promising results of this annotation-projection technique for part-of-speech tagging, named entities, and morphology, and at Maryland we have been focusing on the challenges of projecting syntactic dependency relations.

Figure 1 shows our baseline architecture, which includes not only the creation of a noisy treebank but also its application in an end-to-end machine translation process. Briefly, a word-aligned parallel corpus is created as discussed in the previous section. The English side is analyzed using Dekang Lin's Minipar parser (Lin, 1997), which produces syntactic dependencies, e.g. indicating arguments of verbs, modifiers, etc. Crucially, the resulting dependency representation is independent of word order.

Projection of syntactic dependencies relies on a fairly strong hypothesis: that major grammatical relations are preserved across languages. Operationally, the transfer process begins by assuming that if words e_1 and e_2 in English correspond to s_1 and s_2 in Spanish, respectively, and there is a dependency relation r between e_1 and e_2,

then r will hold between s_1 and s_2. For example, 'black cat' in English corresponds to 'gato negro' in Spanish. Therefore the relationship adjmod(cat,black) is transferred into the Spanish analysis as adjmod(gato,negro). Notice that the relationship abstracts away from word order. These resulting representations constitute a noisy dependency treebank, which we are using as the training set for Ratnaparkhi's (1997) MXPOST POS tagger and Collins's (1997) stochastic parser.

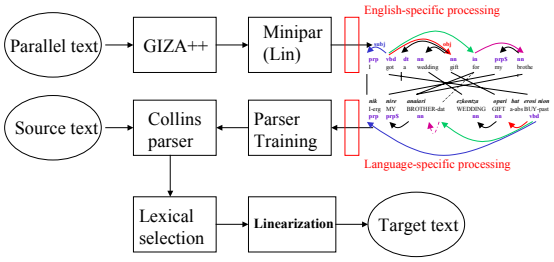

Figure 1. Baseline Dependency Transfer Architecture

As stated, the hypothesis of direct dependency transfer is clearly false – indeed, the issue of divergences in translation has been an important focus in our previous work (Dorr, 1993). However, we are optimistic that cross-language correspondence of dependencies is a suitable starting point for investigation on both theoretical and empirical grounds. Theoretically, grammatical relations are closer than constituency relations to the thematic relationships underlying the sentence meaning common to both sides of the translation pair; thus the fundamental correspondences are likely to hold much of the time. Moreover, lexical dependencies have proven to be instrumental in advances in monolingual syntactic analysis (e.g.

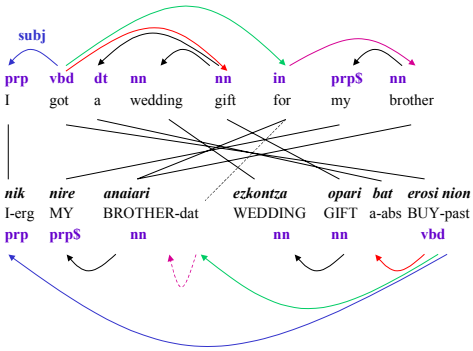

Figure 2. An example of dependency transfer

Collins, 1997). These considerations distinguish our approach from Wu's (2000) approach, which characterizes the cross-language syntactic relationships using a non-lexicalized bilingual grammar formalism.

Our second cause for optimism is empirical: in preliminary efforts we have attempted the direct dependency transfer approach with Spanish and Chinese, with bilingual speakers and linguists inspecting the results. The results of dependency transfer look promising, and the problems that are evident so far tend to be linguistically interesting and amenable to language-specific post-transfer processing. As one example, English parses projected into Spanish will not lead to useful dependencies involving the reflexive *se* when, as is often the case, it has no lexically realized correspondent on the English side; post-processing of the Spanish can be used to introduce a dependency relationship between the verb and the reflexive morpheme. The use of English-side information contrasts with the unsupervised dependency-based translation models of Alshawi *et al.* (2000).

Figure 2 provides an illustrative example using English and Basque, which have very different linguistic properties. The figure shows that the verb-subject, verb-object, and modification relationships (most dependency labels suppressed) transfer directly to the Basque sentence (a fluent translation in neutral word order). The indirect object relationship is expressed in the English parse via prepositional modification between 'got' and 'for', together with the relationship between 'for' and 'brother'; on the Basque side the dative component of meaning and the morpheme for 'brother' are conflated in the word 'anaiari'; the resulting pattern of syntactic dependency links on the Basque side can be post-processed, with the word-internal dependency being converted into a lexical feature.

As an important part of our initial efforts, we are developing rigorous evaluation criteria based on precision and recall of dependency triples, using manually created dependencies as a gold standard and using inter-annotator precision and recall to provide an upper bound.

Improving Quality in Broad-Coverage MT

Analysis and evaluation of MT output from existing systems (including Systran) reveals that there is a great deal of work to be done to provide improved

quality. We are currently focusing our efforts on (a) providing linguistically motivated knowledge to enhance our existing source-language parsing module; (b) using additional knowledge about divergence categories to improve on alignments between source- and target-language dependencies; and (c) conditioning statistical translation components, including parsing to and generation from dependency structures, on linguistic features not currently taken advantage of in the traditional IBM-style models.

As one example, we take advantage of semantically classed verbs (Dorr, 1997) to capture valence and other linguistic information to improve parsing operations such as PP attachment. For example, all verbs in the class {arrange, immerse, install, lodge, mount, place, position, put, set, situation, sling, stash, stow} take a locative prepositional phrase as an argument; if our training data contains only the most frequently occurring verbs in this class (such as `put'), we can deduce, by association, that others (such as `sling') have the same PP attachment properties – and thus can improve parsing for these sparsely occurring verbs.

As another example, stochastic alignment algorithms are likely to map the English predicate `kick' to the corresponding French `coup' (especially since the two words also co-occur as nouns in the absence of `donner' leaving the actual predicate `donner' unaligned when we generate the aligned dependency-tree database (to be described in the next section). This just one instance of a more general phenomenon: languages sometimes package up elements of meaning, particularly verb meaning, into different constituents than English does (i.e., language divergences). To address this issue, we pre-process the English using the semantically classed verbs, so that we automatically expand verbs in selected divergence classes into alignable constituents.

A third example is the use of supervised word sense disambiguation techniques in lexical selection. We have developed a set of tools for supervised WSD that uses a combination of broad-window and local collocational features to represent contexts for an ambiguous word. A variety of classification algorithms can be used – we obtained promising results for English, Spanish, and Swedish in the recent SENSEVAL-2 evaluation exercise (Cotton et al., 2001) using support vector machines for the classification process. We are currently investigating the adaptation of this method to perform lexical selection, with the target English word playing the same role as the sense tag for Spanish words.

References

H. Alshawi, B. Srinivas, and S. Douglas. 2000. Learning dependency translation models as collections of finite state head transducers. *Computational Linguistics*, 26.

Al-Onaizan *et al*. 1999., Statistical MT: Final Report, Johns Hopkins Summer Workshop 1999, CLSP Technical report, Johns Hopkins University.

Michael Collins. 1997. Three generative, lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the ACL,* Madrid, Spain.

S. Cotton, P. Edmonds, A. Kilgarriff, and M. Palmer. 2001. SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems, Toulouse, July 2001.

Bonnie J. Dorr. 1993. Machine Translation: A View from the Lexicon. The MIT Press, Cambridge, MA.

Bonnie J. Dorr. 1997. Large-Scale Acquisition of LCS-based lexicons for foreign language tutoring. In *Proceedings of the ACL 5th Conference on Applied Natura lLanguage Processing* (ANLP), Washington, DC.

Bonnie J. Dorr and Douglas W. Oard. 1998. Evaluating resources for query translation in cross-language information retrieval. In *1st International Conference on Language Resource Evaluation* (LREC), Granada, Spain.

Ray Jackendoff. 1983. Semantics and Cognition. The MIT Press, Cambridge, MA.

Ray Jackendoff. 1990. Semantic Structures. The MIT Press, Cambridge, MA.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th ACL/EACL'97*, Madrid, Spain, July.

Adwait Ratnaparkhi. 1996. A Maximum entropy model for part-of-speech tagging. In Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania.

Philip Resnik. 1999. Mining the web for bilingual text. In *37th Annual Meeting of the Association for Computational Linguistics* (ACL'99), College Park, Maryland, June.

Philip Resnik, Douglas Oard, and Gina Levow. 2001. Improved cross language retrieval using backoff translation. 2001. *Proc. HLT'2001*. San Diego.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 2000. The Bible as a Parallel Corpus: Annotating 'The Book of 2000 Tongues'. Computers and the Humanities, 33(1-2).

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A Maximum entropy approach to identifying sentence boundaries. *Proceedings of the ACL 5th Conference on Applied Natural Language Processing* (ANLP), Washington, DC.

Dekai Wu. 2000. Alignment using stochastic inversion transduction grammars. In Jean Veronis, editor, *Parallel Text Processing*. Kluwer.

David Yarowsky, Grace Ngai, and Richard Wicentowsky,. 2001. Inducing Multilingual Text and Tools via Robust Projection Across Aligned Corpora. Proc. *HLT'2001*, San Diego.